

# Assessing the stability of a phylogeny without bootstrap: an analytical approach

Mahendra Mariadassou

June 19, 2007

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Framework</b>	<b>3</b>
2.1	Notations and definitions . . . . .	3
2.2	Toy Example . . . . .	4
2.3	Likelihood and stability . . . . .	5
<b>3</b>	<b>Phylogenetic reconstruction for finite size samples</b>	<b>7</b>
3.1	Connection between $\ell^T$ and $Q$ . . . . .	8
3.2	Distance between $Q$ and $Q_n$ . . . . .	8
3.3	Stability of the inferred tree . . . . .	11
<b>4</b>	<b>Comparison with widely used methods</b>	<b>13</b>
4.1	Bootstrap . . . . .	13
4.2	Comparison on the toy example . . . . .	14
<b>A</b>	<b>Proofs of the preliminary results</b>	<b>17</b>
A.1	Proof of the first result . . . . .	17
A.2	Proof of the second result . . . . .	19
	<b>References</b>	<b>20</b>

# 1 Introduction

Phylogenies, or evolutionary trees, are the basic structures necessary to analyze differences between species and to analyze these differences statistically. Several methods are available to infer phylogenies, the two most popular being Maximum Parsimony (MP) and Maximum Likelihood (ML) estimation (see [11] for a comprehensive review). The ML method [6] provides a statistical framework to answer this question, whereas the MP method does not. We therefore focus on ML methods.

We are interested in the stability of the inferred phylogeny. Several bootstrap methods have been developed to specifically address this issue (see [3], [4], [2], [14] and [7] for a review). Stability is a desirable, if not necessary, property for a phylogeny: after inferring a tree, we want to draw some conclusions from it. Since we draw general conclusions, we want the tree to be as stable as possible: a small modification in the data should not drastically change the phylogeny and invalidate our conclusions, or at least if it does it should only do so with a small probability. An inferred phylogeny not holding this property has a very limited utility: no conclusions drawn from it would be useful.

A most common stability problem, from which bootstrap in phylogeny originates (see [3], [13]), is the support given to a clade. For example if a phylogeny classes species  $A$  in a different clade than species  $B$  and  $C$ , we want to assess the significance of this classification: is the clade supported by a lot of evidence or is it just here by chance ?

Most bootstrap methods are based on re-sampling with replacement [1] and account only for the observed variability. Bootstrap methods also discard the relation between the size of the data, the number of species in the study and the stability of the phylogeny.

In this paper we propose an analytical approach to this issue. Rather than working on the phylogeny, we work on their likelihood score: stable likelihood scores are equivalent to a stable ranking and thus to a stable ML phylogeny. We obtain bounds on the probability that the empirical likelihood wanders too far away from its average value. One obvious advantage of doing so is to reduce the study of phylogenies and phylogenetic trees to the much simpler study of likelihood scores taking values in  $\mathbb{R}$ .

Section 2 is devoted to the framework we use. We also introduce the notations and illustrate them on a toy example. Then, in Section 3, we present our main result concerning the stability of the likelihood score. Finally in Section 4, we compare our method to those used in the literature and discuss our results. Technical proofs of some results are postponed in appendix A.

## 2 Framework

We introduce here the statistical framework in which we work and the notations we use and illustrate them on a toy example.

### 2.1 Notations and definitions

The data set in our analysis is a  $s \times n$  matrix  $\mathcal{X} = (X_1, \dots, X_n) = (X_{ij})_{i=1\dots n, j=1\dots s}$  representing a set of molecular sequences aligned over different species.  $n$  is the length of the sequence after the alignment (including gaps) and  $s$  the number of species.  $X_{ij}$  takes value in an alphabet  $\mathcal{A}$  and codes for the state of the  $i$ th nucleotide of the alignment in the  $j$ th species. The  $j$ th line of  $\mathcal{X}$  is then the aligned sequence corresponding to species  $j$ . When working with DNA sequences,  $\mathcal{A}$  is usually  $\{A, C, G, T\}$  but it can take others values, for example when working with protein sequences. The statistical unit of interest is the column  $X_i$ , a  $s$ -dimensional vector valued in  $\mathcal{A}^s$ , which codes for the pattern of nucleotide  $i$  over all  $s$  species.

We assume that the pattern  $X_i$  at  $i$ th site is an observed value of random variable  $X$  and that the  $X_i$ s are *i.i.d.* The probability function of  $X$  is  $Q$ .

**Definition 1.** We define a phylogenetic model, or phylogeny,  $T$  as the union of:

- (i) the evolution model: the substitution model and its associated parameters,
- (ii) the tree topology: a branching pattern and the associated branch lengths

Under model  $T$ , the log-likelihood of an observed pattern  $x$  (or equivalently of a site presenting this pattern) is:

$$\ell^T(x) := \log \mathbb{P}(x; T)$$

**Definition 2.** The empirical (resp. true) mean log-likelihood  $\ell_n^T$  (resp.  $\ell^T$ ) is the mean of  $\log \mathbb{P}(X; T)$  under the empirical (resp. true) distribution:

$$\ell_n^T = \mathbb{E}_{Q_n}[\log \mathbb{P}(X; T)] = \frac{1}{n} \sum_i \log \mathbb{P}(X_i; T) \quad (1)$$

$$\ell^T = \mathbb{E}_Q[\log \mathbb{P}(X; T)] = \sum_{x \in \mathcal{A}^s} Q(x) \log \mathbb{P}(x; T) \quad (2)$$

with the empirical distribution of the patterns defined as:

$$Q_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$$

The empirical distribution is opposed to the unknown *true distribution*  $Q$  of the patterns, which is unachievable from the data except for a infinite number of nucleotides, when  $n = \infty$ .

We define in the same way, the *empirical* (resp. *true*) *variance*  $\sigma_n^2(T)$  (resp.  $\sigma^2(T)$ ) by:

$$\begin{aligned}\sigma_n^2(T) &= \mathbb{V}_{Q_n}[\log \mathbb{P}(X; T)] = \frac{1}{n} \sum_i (\log \mathbb{P}(X_i; T) - \ell_n^T)^2 \\ \sigma^2(T) &= \mathbb{V}_Q[\log \mathbb{P}(X; T)] = \sum_{x \in \mathcal{A}^s} Q(x) (\log \mathbb{P}(x; T) - \ell^T)^2\end{aligned}$$

We need the expressions of both the empirical and the true log-likelihood and variance. The goal, as presented in details in 3.3 is to compare not only the empirical mean log-likelihood to its true average but also the rankings induced on phylogenetic models by the both the empirical and the true mean log-likelihood.

## 2.2 Toy Example

To illustrate these quantities, consider a toy example made of  $s = 4$  species called  $S_1, S_2, S_3$  and  $S_4$  and  $n = 3$  nucleotides. Although  $n = 3$  is too low to be realistic, this example usefully serves our purpose. The data are a  $4 \times 3$  matrix  $\mathcal{X}$ :

$$\mathcal{X} = \begin{array}{c|cc} & \textit{Species} & \textit{Sites} \\ \hline & S_1 & A \ A \ A \\ & S_2 & G \ G \ C \\ & S_3 & C \ C \ A \\ & S_4 & C \ C \ C \end{array}$$

Note that the first two patterns are identical and require at least a transition and a transversion whereas the third one minimal requirement is only a transversion.

We consider only three phylogenies sharing a *Kimura two-parameter (K2P)* evolution model (see [9]) with transversion rate  $\alpha$  and transition rate  $\beta$  and the following unrooted unlabeled topology pattern:

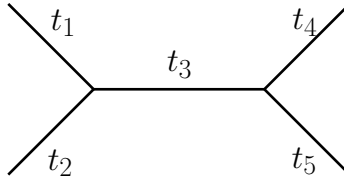


Figure 1: Topology with unequal branch lengths :  $t_1 > t_2$ ,  $t_4 > t_5$  and  $t_3 > t_1, t_2, t_4, t_5$

We chose a  $K2P$  evolution model because it is the simplest realistic model: although the Jukes-Cantor model (see [8]) is even simpler, it is too unrealistic to be interesting. The three different phylogenies are obtained when considering the three possible topologies, each topology corresponding to a different labeling, as shown in Fig. 2.

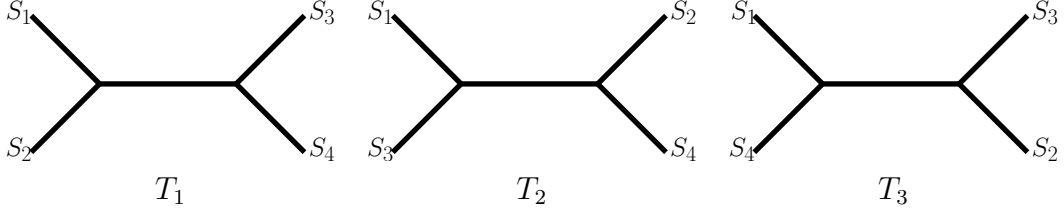


Figure 2: The three topologies obtained by labeling the nodes

We assume that  $\alpha, \beta \ll 1$  so that the  $\alpha t_{i,s}$  and  $\beta t_{i,s}$  are also very small and that in each model only one configuration of the inner nodes contributes to the log-likelihood. In this special case, parsimony and maximum likelihood, the two main methods, agree: since for a fixed topology an evolutionary change occurs only with a very low probability, the most likely configuration is the one needing the less such changes, namely the most parsimonious. All others are far less likely, and as such, barely contribute to the log-likelihood of the topology.

Since the  $X_i$  are i.i.d, we have:

$$\begin{aligned} 3\ell_n^{T_1} &= 2 \log \mathbb{P}(AGCC; T_1) + \log \mathbb{P}(ACAC; T_1) \\ 3\ell_n^{T_2} &= 2 \log \mathbb{P}(AGCC; T_2) + \log \mathbb{P}(ACAC; T_2) \\ 3\ell_n^{T_3} &= 2 \log \mathbb{P}(AGCC; T_3) + \log \mathbb{P}(ACAC; T_3) \end{aligned}$$

To compute these probabilities, we need to specify an ranking on the branch lengths: we adopted  $t_1 > t_2$ ,  $t_4 > t_5$  and  $t_3 > t_1, t_2, t_4, t_5$ . The central branch is the longest and both upward branches are longer than their downward counterpart. Some simple computations taking advantage of the previous remark then give:

### 2.3 Likelihood and stability

The toy example is simple so we can easily calculate probabilities in it. We will start by calculating both the likelihood score and the variance associated to each model, as defined in Sec. 2.1. The results are shown in Tab. 2

At this point, we ask ourselves which model has the best likelihood score and which has the lowest variance. Although the interest of likelihood scores is obvious in a ML estimation framework, the interest of variances is a bit more intricate. A

Table 1: Log-likelihood of the two patterns for each of the three model

Model	Pattern	Likelihood
1	AGCC	$\log \beta t_1 + \log \alpha t_3$
1	ACAC	$\log \beta t_1 + \log \beta t_4$
2	AGCC	$\log \beta t_1 + \log \beta t_4$
2	ACAC	$\log \alpha t_3$
3	AGCC	$\log \beta t_1 + \log \beta t_5$
3	ACAC	$\log \beta t_1 + \log \beta t_4$

Table 2: Log-likelihood of the two patterns for each of the three model.

Model $i$	Likelihood $\ell_n^{T_i}$	Variance $\sigma_n^2(T_i)$
1	$3 \log \beta t_1 + \log \beta t_4 + 2 \log \alpha t_3$	$\frac{2}{9} \log^2 \frac{\alpha t_3}{\beta t_4}$
2	$2 \log \beta t_1 + 2 \log \beta t_4 + \log \alpha t_3$	$\frac{2}{9} \log^2 \frac{\beta^2 t_1 t_4}{\alpha t_3}$
3	$3 \log \beta t_1 + \log \beta t_4 + 2 \log \beta t_5$	$\frac{2}{9} \log^2 \frac{t_5}{t_4}$

low variance means the likelihood of a given nucleotide is close to the likelihood of the model so that all nucleotides almost evenly support the tree. By opposition a high variance means nucleotides are discordant in their support: some of them strongly support the model whereas others only have a weak support. In a high variance context, adding or removing a nucleotide from the data may dramatically change the empirical likelihood of the model. And in particular the empirical likelihood may be quite different from the true one. A low variance is thus a desirable property.

**Proposition 3.** *Although  $T_1$  is the natural choice, under conditions:*

$$\alpha t_1 t_3 < t_4 \quad \text{and} \quad \beta t_1 t_5^2 < \alpha t_3 t_4 \quad (3)$$

$$\left| \frac{t_5}{t_4} - 1 \right| < \left| \frac{\alpha t_3}{\beta t_4} - 1 \right| \quad \text{and} \quad \left| \frac{t_5}{t_4} - 1 \right| < \left| \frac{\beta^2 t_1 t_4}{\alpha t_3} - 1 \right|, \quad (4)$$

$T_3$  is the lowest variance model whereas  $T_2$  is the highest likelihood score model.

Let us discuss the scope of this results before turning to its proof.

**Remark:** since  $X_1$  and  $X_2$  support model  $T_1$ , it is a natural choice but it has neither the best likelihood score nor the lowest variance.  $T_2$  is strongly supported

by nucleotide  $X_3$  and weakly by the the others. As such  $T_2$  is the ML-model for these 3 nucleotides but it may lose this title when adding or removing a nucleotide: its stability is bad. On the other hand  $T_3$  has a low variance, close to 0: none of the nucleotides support it so that the likelihood of each is uniformly bad and variance among nucleotides is very small. Prop. 3 proves that neither likelihood score nor variance alone are good criteria.

**Remark:** note that conditions given in (3) and (4) are very mild. In fact as soon as the branch lengths are of the same order of magnitude, three of them are consistent with and immediately satisfied under the assumption made in Sec. 2.2 that the  $\alpha t_i$ s and  $\beta t_i$ s are very small. The remaining one is:  $|\frac{t_5}{t_4} - 1| - |\frac{\alpha t_3}{\beta t_4} - 1|$ , which is still very mild. Our branching order specifies  $t_4 > t_5$  and  $t_3 > t_4$ , so it boils down to  $\frac{\alpha t_3}{\beta t_4} > \frac{t_4}{t_5}$ . Since in an ordinary  $K2P$  model,  $\alpha > \beta$ , this condition is satisfied for a wide scope of  $t_3, t_4$  and  $t_5$ .

**Proof.** The condition for  $T_2$  to be the ML-model is  $\ell_n^{T_2} \geq \max(\ell_n^{T_1}, \ell_n^{T_3})$  which, given our branch lengths ranking and the values given in Tab 2, boils down to:

$$\begin{aligned}\ell_n^{T_2} - \ell_n^{T_1} &= \log \frac{t_4}{\alpha t_1 t_3} > 0 \\ \ell_n^{T_2} - \ell_n^{T_3} &= \log \frac{\alpha t_3 t_4}{\beta t_1 t_5^2} > 0\end{aligned}$$

which is true under conditions (3)

The condition for  $T_3$  to be the lowest variance model is  $\sigma_n^2(T_3) \geq \min(\sigma_n^2(T_1), \sigma_n^2(T_2))$  which, given our branch lengths ranking and the values given in Tab 2, boils down to:

$$\begin{aligned}\sigma_n^2(T_3) - \sigma_n^2(T_1) &= \frac{2}{9} \left( \log^2 \frac{t_5}{t_4} - \log^2 \frac{\alpha t_3}{\beta t_4} \right) < 0 \\ \sigma_n^2(T_3) - \sigma_n^2(T_2) &= \frac{2}{9} \left( \log^2 \frac{t_5}{t_4} - \log^2 \frac{\beta^2 t_1 t_4}{\alpha t_3} \right) < 0\end{aligned}$$

which is true under conditions (4) ■

### 3 Phylogenetic reconstruction for finite size samples

In this section, we study the behavior of  $\ell_n^T$  with respect to  $\ell^T$ , especially the effect of  $n$  on the distance between  $\ell_n^T$  and  $\ell^T$ . We do this in three steps: we first connect  $\ell_n^T$  and  $\ell^T$  to  $Q_n$  and  $Q$ , then calculate some bounds on  $Q_n - Q$  and finally adapt these bounds to  $\ell_n^T - \ell^T$ .

### 3.1 Connection between $\ell^T$ and $Q$

We start by defining some vectors  $\boldsymbol{\theta}$  and  $\boldsymbol{\theta}_n$  coding for the same information as the distributions  $Q$  and  $Q_n$ . We do because vectors are easier to manipulate than distributions.

**Definition 4.** *The distributions  $Q$  and  $Q_n$  introduced in Sec. 2.1 are defined over  $\mathcal{A}^s$ . As such, they are completely by the vectors  $\boldsymbol{\theta} = (\theta^x)_{x \in \mathcal{A}^s}$  and  $\boldsymbol{\theta}_n = (\theta_n^x)_{x \in \mathcal{A}^s}$  of length  $|\mathcal{A}^s|$ , where for all  $x \in \mathcal{A}^s$ :*

$$\begin{aligned}\theta^x &= \mathbb{P}_Q(X = x) \\ \theta_n^x &= \mathbb{P}_{Q_n}(X = x)\end{aligned}$$

With  $\ell^T$  defined as in Def. 2, classical properties of the *Kullback – Leibler* distance (KL-distance) ensures that maximizing  $\ell^T$  over models  $T$  is equivalent to minimizing the KL-distance between  $\mathbb{P}(\cdot; T)$ , the pattern distribution under model  $T$ , and  $Q$ , its true distribution.

Just like  $X$ ,  $\log \mathbb{P}(\cdot; T)$  is discrete-valued. We therefore consider the vector  $\log P^T$  of size  $|\mathcal{A}^s|$  defined by  $\log P^T = (\log \mathbb{P}(x, T))_{x \in \mathcal{A}^s}$ .

Using previous notations

$$\ell^T = \ell_n^T + (\boldsymbol{\theta} - \boldsymbol{\theta}_n)' \cdot \log P^T \quad (5)$$

This result simply comes from:

$$\begin{aligned}\ell^T &= \mathbb{E}_Q[\log \mathbb{P}(X; T)] = \sum_{x \in \mathcal{A}^s} \theta^x \log \mathbb{P}(x; T) \\ \ell_n^T &= \mathbb{E}_{Q_n}[\log \mathbb{P}(X; T)] = \sum_{x \in \mathcal{A}^s} \theta_n^x \log \mathbb{P}(x; T)\end{aligned}$$

The true log-likelihood  $\ell^T$  is the sum of two quantities, the computable *empirical* log-likelihood  $\ell_n^T$  and the unknown correction term  $(\boldsymbol{\theta} - \boldsymbol{\theta}_n)' \cdot \log P^T$ . To control the difference  $\ell^T - \ell_n^T$ , the model  $T$  is not enough, we also need information on the difference  $\boldsymbol{\theta} - \boldsymbol{\theta}_n$ .

### 3.2 Distance between $Q$ and $Q_n$

As suggested by (5), we now focus on  $(\boldsymbol{\theta} - \boldsymbol{\theta}_n)$ . Our goal is to bound the probability of it being "large". Since  $\boldsymbol{\theta} - \boldsymbol{\theta}_n$  is a  $|\mathcal{A}^s|$ -dimension random variable, we proceed in three steps: we first turn our question into a series of unidimensional questions, then work on each of them separately and finally combine them before turning back to the original question.

**Simplification of the question**

**Lemma 5.** Let  $\boldsymbol{\theta} = (\theta^x)_{x \in \mathcal{A}^s}$  and  $\|\boldsymbol{\theta}\|_\infty = \max_{x \in \mathcal{A}^s} |\theta^x|$ , then:

$$\mathbb{P}(\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| > \varepsilon) \leq |\mathcal{A}|^s \max_{x \in \mathcal{A}^s} \mathbb{P}(|\theta^x - \theta_n^x| > \varepsilon)$$

The cardinal of the possible states space, here  $|\mathcal{A}|^s$  play a crucial in the formula as a multiplicative factor of the probability.

**Proof.** The results comes from

$$\begin{aligned} \mathbb{P}(\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\|_\infty > \varepsilon) &= \mathbb{P}\left(\max_x |\theta^x - \theta_n^x| > \varepsilon\right) \\ &= \mathbb{P}\left(\bigcup_x \{|\theta^x - \theta_n^x| > \varepsilon\}\right) \\ &\leq \sum_x \mathbb{P}(|\theta^x - \theta_n^x| > \varepsilon) \\ &\leq |\mathcal{A}|^s \max_{x \in \mathcal{A}^s} \mathbb{P}(|\theta^x - \theta_n^x| > \varepsilon) \end{aligned} \quad (6)$$

■

Note first that each of the  $\theta^i - \theta_n^i$  is real-valued. Since  $\theta_n^x - \theta^x = \frac{1}{n} \sum_{j=1}^n (\mathbb{1}_{\{X_j=x\}} - \theta^x)$  and the  $\mathbb{1}_{\{X_j=x\}}$  are i.i.d Bernoulli variables with parameter  $\theta^x$ , we can use standard concentration inequalities to control  $\theta_n^x - \theta^x$ .

**Bernoulli lemma** Forget for now the  $\theta_n^x$  and consider  $(X_n)$  a sequence of i.i.d. random variables with parameter  $p$ . To upper-bound the probability of the event  $\{|\sum_{i=1}^n (X_i - p)| > n\varepsilon\}$ , we need to upper-bound the probabilities of both  $\{\sum_{i=1}^n (X_i - p) > n\varepsilon\}$  and  $\{\sum_{i=1}^n (X_i - p) < -n\varepsilon\}$ .

**Lemma 6.** Consider  $(X_n)$  a sequence of i.i.d. Bernoulli variables with parameter  $p$ , we have:

$$\frac{\log P(\sum_{i=1}^n (X_i - p) > \varepsilon)}{n} \leq \frac{-\varepsilon^2}{2p(1-p-\varepsilon)} \left[1 - \frac{\varepsilon}{p(1-p-\varepsilon)^2}\right] \quad (7)$$

**Lemma 7.** Consider  $(X_n)$  a sequence of i.i.d. Bernoulli variables with parameter  $p$ , we have:

$$\frac{\log P(\sum_{i=1}^n (X_i - p) < -\varepsilon)}{n} \leq \frac{-\varepsilon^2}{2p(1-p+\varepsilon)} \left[1 - \frac{2\varepsilon}{p(1-p+\varepsilon)}\right] \quad (8)$$

We expect the probability of  $\{|\sum_{i=1}^n (X_i - p)| > n\varepsilon\}$  to decrease to 0 with  $n$ . According to large deviations theory, the decay is exponential. The main purpose of these two results is to uncover this exponential speed.

We refer the reader to the appendix for the proofs of Lemma. 6 and 7. Although the two results look alike, one concerns the left end-tail of the distribution of  $\sum_{i=1}^n (X_i - p)$  whereas the other one concerns the right end-tail. We can not deal with them in exactly the same way.

### Concentration result

**Proposition 8.** *For small  $\varepsilon/p$ , when neglecting second order terms, we have:*

$$\frac{\log \mathbb{P}(\|\theta - \theta_n\| > \varepsilon)}{n} \leq \frac{s}{n} \log |\mathcal{A}| + \frac{\log 2}{n} + \max_{x \in \mathcal{A}^s} \frac{-\varepsilon^2}{\theta^x (1 - \theta^x + \varepsilon)} \quad (9)$$

**Proof.** Since

$$\begin{aligned} & \mathbb{P} \left( \left| \frac{1}{n} \sum_{i=1}^n X_i - p \right| > \varepsilon \right) \\ &= \mathbb{P} \left( \frac{1}{n} \sum_{i=1}^n (X_i - p) > \varepsilon \right) + \mathbb{P} \left( \frac{1}{n} \sum_{i=1}^n (X_i - p) < -\varepsilon \right) \\ &:= P_n^1 + P_n^2, \end{aligned}$$

we find:

$$\frac{\log P(|\frac{1}{n} \sum_{i=1}^n (X_i - p)| > \varepsilon)}{n} \leq \frac{\log 2}{n} + \max \left( \frac{\log P_n^1}{n}, \frac{\log P_n^2}{n} \right)$$

We can upper bound  $P_n^1$  using (7) and  $P_n^2$  using (8). When neglecting second order terms in  $\varepsilon/p$  and since  $\frac{-\varepsilon^2}{2p(1-p+\varepsilon)} \geq \frac{-\varepsilon^2}{2p(1-p-\varepsilon)}$  for all  $\varepsilon$ , we get:

$$\frac{\log P(|\frac{1}{n} \sum_{i=1}^n (X_i - p)| > \varepsilon)}{n} \leq \frac{-\varepsilon^2}{2p(1-p+\varepsilon)} + \frac{\log 2}{n}$$

Applying it to (6) gives the result. ■

The  $\theta^x$  giving the smallest decreasing rate for the exponential bound are those close to  $(1 + \varepsilon)/2$ . For this  $\theta^x$ , we can give the explicit worst rate. But since there is no evidence than a pattern should appear more than half the time, we expect the  $\theta^x$  to be significantly smaller than  $(1 + \varepsilon)/2$  and the decreasing rate to be significantly higher.

### 3.3 Stability of the inferred tree

In the following, we take advantage of the previous results on  $\theta - \theta_n$  to bound the difference  $\ell^T - \ell_n^T$ . After doing so, we focus on inversions probabilities, *i.e.* incongruities between the empirical likelihood ranking and the true one.

**Distance between the empirical and true mean log-likelihood** In this part, the goal is to evaluate how confident we are in the log-likelihood given to a tree. The less confident we are, the more cautious we must be when dealing with that log-likelihood, for example when comparing it for two trees. To do so, we connect  $\ell^T - \ell_n^T$  to  $\theta - \theta_n$  using (5).

**Corollary 9.** Note  $\tilde{\varepsilon} = \frac{\varepsilon}{|\mathcal{A}|^s \|\log P^T\|_\infty}$ . Then:

$$\frac{\log \mathbb{P}(|\ell^T - \ell_n^T| \geq \varepsilon)}{n} \leq \frac{s}{n} \log |\mathcal{A}| + \frac{\log 2}{n} + \max_{x \in \mathcal{A}^s} \frac{-\tilde{\varepsilon}^2}{\theta^x (1 - \theta^x + \tilde{\varepsilon})} \quad (10)$$

**Proof.** The result comes from the simple inequality:

$$|\ell^T - \ell_n^T| = |(\theta_n - \theta)' \cdot \log P^T| \leq |\mathcal{A}|^s \|\theta_n - \theta\|_\infty \|\log P^T\|_\infty$$

■

$|\mathcal{A}|^s$  is the theoretic number of possible patterns and appears twice in the formula, once as a multiplicative factor  $s \log |\mathcal{A}|$  and once as a rescaling of the rescaling deviation  $\varepsilon$ . Replacing  $|\mathcal{A}|^s$  by a smaller value gives a finer bound. For example, dividing it by 2 gives a multiplicative factor twice smaller and a decreasing rate at least four times bigger.

The patterns  $x$  appearing with probability  $\theta^x = 0$  also satisfy  $\theta_n^x$ : remember that  $\theta_n^x$  is the apparition frequency of  $x$  in a multinomial  $n$ -sample with parameter  $\theta$ . As such, they do not contribute in the smallest to  $|(\theta_n - \theta)' \cdot \log P^T|$  and can be removed without changing the results. We can therefore replace  $|\mathcal{A}|^s$  by the true number  $N_0$  of possible patterns and  $\|\log P^T\|_\infty$  by  $\|\log P^T\|_{app} = \max_{\substack{x \in \mathcal{A}^s \\ \theta^x > 0}} |\log P(x; T)|$ .

Although  $N_0$  is unknown, we expect it to be significantly smaller than  $|\mathcal{A}|^s$  so that even a roughly overestimated  $\hat{N}_0$  will tremendously increase our bound and the decreasing rate.

Prop. 9 also hints as to how  $n$  should evolve with  $s$ . As the number of species  $s$  increases, so do  $N_0$  and  $\|\log P^T\|_{app}$ . The multiplicative factor grows while the

decreasing rate shrinks. To maintain the probability of  $\{|\ell_n^T - \ell^T| > \varepsilon\}$  low, we must increase the sample size  $n$  to compensate for both these effects. For example, in the simple case where  $Q$  and  $P(\cdot; T)$  are uniform over all  $\{A, C, G, T\}^s$ , (10) turns out to be:

$$\log \mathbb{P} (|\ell^T - \ell_n^T| \geq \varepsilon) \leq s \log 4 + \log 2 - n \frac{\varepsilon^2}{4^s s^2 \log^2 4}$$

To maintain a given level of confidence when adding a new species (that is replacing  $s$  by  $s + 1$ ), we should replace  $n$  by  $4n(1 + 1/s^2) + (s + 1)^2 4^{s+1} \log^3 4 / \varepsilon^2$ . At first sight, the cost of an additional species is prohibitive, but remember we considered the worst case (all  $|\mathcal{A}|^s$  patterns are possible) for educational purpose. In general, the cost of an additional species will be considerably lower than that. Also remark that when including one or more additional species in the study, we change the model  $T$  of interest. In particular, the topology embedded in model  $T$  no longer has only  $s$  leaves, but more than  $s$ .

**Support given to a tree** A further approach is to focus on the ranking on the models induced by their mean log-likelihood. Since we do not have access to the true log-likelihoods of models, we base on our ranking on their empirical log-likelihoods. Of course, inversion events can happen: when comparing two models  $T$  and  $T'$ , the empirical log-likelihoods could give a different ranking than the true ones. Since we retrieve the maximum *empirical* log-likelihood model, this is an unwanted event. We offer here to bound the probability of such an event as a function of the relevant quantities.

**Proposition 10.** *Assume that tree  $T$  is better than tree  $T'$  in the sense that  $\ell^T > \ell^{T'}$ . Then, the probability  $P_n$  that  $T'$  is better than  $T$  for our sample:  $P_n = \mathbb{P}(\ell_n^T - \ell_n^{T'} < 0)$  is such that:*

$$\frac{\log \mathbb{P}(\ell_n^T - \ell_n^{T'} < 0)}{n} \leq \frac{s}{n} \log |\mathcal{A}| + \max_{x \in \mathcal{A}^s} \frac{-\varepsilon^2}{\theta^x (1 - \theta^x + \varepsilon)} \quad (11)$$

where  $\varepsilon = \frac{\ell^T - \ell^{T'}}{|\mathcal{A}^s| \|\log P^T - \log P^{T'}\|_\infty}$

**Proof.** Since  $\ell_n^T - \ell_n^{T'} = \frac{1}{n} \sum_{i=1}^n \log \mathbb{P}(X_i; T) - \log \mathbb{P}(X_i; T')$ , we can use (7) and (8) to bound  $\ell_n^T - \ell_n^{T'} - (\ell^T - \ell^{T'})$  in the same way than  $\ell_n^T - \ell^T$ . We just need to

replace  $\|\log P^T\|_\infty$  by  $\|\log P^T - \log P^{T'}\|_\infty$ .

$$\begin{aligned}
\Delta &= \mathbb{P}\left(\ell_n^T - \ell_n^{T'} < 0\right) \\
&= \mathbb{P}\left(\ell_n^T - \ell_n^{T'} - (\ell^T - \ell^{T'}) < -(\ell^T - \ell^{T'})\right) \\
&\leq \mathbb{P}\left(\bigcup_{x \in \mathcal{A}^s} (\theta_n^x - \theta^x) \log \frac{\mathbb{P}(x; T)}{\mathbb{P}(x; T')} < -\frac{\ell^T - \ell^{T'}}{|\mathcal{A}|^s}\right) \\
&\leq |\mathcal{A}|^s \max_{x \in \mathcal{A}^s} \mathbb{P}\left(\theta_n^x - \theta^x \leq -\frac{\ell_n^T - \ell_n^{T'}}{\|\log P^T - \log P^{T'}\|_\infty |\mathcal{A}|^s}\right) \tag{12}
\end{aligned}$$

Since we consider only one-sided deviations, combining (12) with (8) gives the result. ■

This result is comforting: the farther  $\ell^T$  and  $\ell^{T'}$  are, the more unlikely an inversion event between  $T$  and  $T'$  is. The comments made for Cor. 9 still hold for Prop. 10. In short, we should replace  $|\mathcal{A}|^s$  by the real number  $N_0$  of patterns and the maximum  $\|\log P^T - \log P^{T'}\|_\infty$  over all patterns by the same maximum  $\|\log P^T - \log P^{T'}\|_{app}$  but only over those patterns we observe.

## 4 Comparison with widely used methods

Several methods already exist in the literature to test the stability of a tree. The most popular is the bootstrap, widely used to give a confidence index to each clade in the tree. Our concentration inequalities basically have the same goal: control the fluctuations of a random estimator around its asymptotic value, but we are interested in the mean likelihood of nucleotide under model  $T$  rather than in the tree embedded in model  $T$ .

### 4.1 Bootstrap

Bootstrap procedures in phylogeny are based on sampling with replacement in the data (see [7]). For example, to test the stability of a clade present in the tree  $T^{(0)}$  inferred from the data, we draw  $B$  bootstrap samples  $(\mathcal{X}^{(i)})_{i=1..B}$  from the data and infer the Maximum-Likelihood tree  $T^{(i)}$  for each sample. Note  $b$  the number of  $T^{(i)}$  in which the clade of interest is present. The bootstrap support to our clade is then  $(b + 1)/(B + 1)$ ; we add the original sample to the  $B$  bootstrap samples. This support is compared to an arbitrary threshold, usually 0.66 or 0.95, and if it is higher, the clade is declared present with a 0.66 or 0.95 support.

On top of the already known and widely discussed problems of bootstrap (see [2], [12], [7] and [5], chapter 20) we claim that setting the threshold *ex ante* has

some disadvantages. In our opinion, the major one is to leave out both  $n$  and  $s$ . On the one hand, for small  $n$ , the stochastic effects can be large so that even if the clade is truly present, it can be absent from bootstrap trees more often than 5% of the times. On the other hand, when  $n$  is large enough, the inferred tree has a high probability of having the same topology and thus the same clades as the true one. In this case, a clade present truly present in the tree will be appear in more than 95% of the bootstrap trees. It is then interesting to set threshold higher than 0.95 to build a more conservative test. In any case, the threshold should be decided with regards both to the number  $s$  of species and  $n$  of nucleotides.

As stated in 3.3, to achieve a given level of confidence on a phylogenetic model over  $s$  species, we need a certain number  $n$  of nucleotides. Of course this number grows, possibly quite rapidly, with  $s$ . While bootstrap procedures are powerless to calculate this  $n$ , we can retrieve it using our analytical techniques. Analytical bounds are also very comfortable in the sense that they let study us both convergence speeds and the importance of initial hypothesis on the stability of the mean log-likelihood.

Finally bootstrap has some limitations as it relies on simulations to compute support probability. For large value of  $n$  and  $s$ , the computational burden can be prohibitive. Our method upper bounds such probabilities instead of approximating them, but in an analytical way: the computational burden is not such a problem anymore.

## 4.2 Comparison on the toy example

**Bootstrap** We want to compare bootstrap with our method on the toy example. Since there are only 3 nucleotides, we can explicitly compute the distribution of the bootstrap samples and bootstrap probability of the event  $\{\ell_n^{T_1} > \ell_n^{T_2}\}$ , namely " $T_1$  is better than  $T_2$ ". Since the order does not matter for the configuration and there are only two patterns in the data ( $X_1 = X_2$ ), the bootstrap samples are  $\{(0, 3), (1, 2), (2, 1), (3, 0)\}$  where the first term in the couple is the number of  $X_1$ s and the second one is the number of  $X_3$ s.  $X_1$  and  $X_3$  are drawn with respective probabilities  $2/3$  and  $1/3$ . The distribution of the bootstrap samples is given in Tab. 3. For each of the bootstrap sample, we compute the likelihood of models  $T_1$  to  $T_3$  and the ranking induced induced by these likelihoods. The results are presented in Tab. 4. Finally, the bootstrap probability of  $\{\ell_n^{T_2} > \ell_n^{T_1}\}$  is  $\mathbb{P}_{boot}(\ell_n^{T_2} > \ell_n^{T_1}) = 19/27$ .

**Analytical Bounds** We want to upper-bound the same probability using our method. We do so using Prop. 10. The only undecided parameter is the true number of patterns. It can be as large as  $4^4 = 256$ , as low as the number of

Table 3: Probabilities of each of the four possible bootstrap samples.

Sample	Probability
(0, 3)	1/27
(1, 2)	2/9
(2, 1)	4/9
(3, 0)	8/27

Table 4: Likelihood score of the 3 models and induced ranking for each of the 4 bootstrap samples.

Sample	Likelihood score of model			Ranking
	$T_1$	$T_2$	$T_3$	
(0, 3)	$3 \log \beta t_1 + 3 \log \beta t_1$	$3 \log \alpha t_3$	$3 \log \beta t_1 + 3 \log \beta t_4$	$T_2 > T_1 = T_3$
(1, 2)	$3 \log \beta t_1 + 2 \log \beta t_4$ $+ \log \alpha t_3$	$\log \beta t_1 + \log \beta t_4$ $+ 2 \log \alpha t_3$	$3 \log \beta t_1 + 2 \log \beta t_4$ $+ \log \beta t_5$	$T_2 > T_1 > T_3$
(2, 1)	$3 \log \beta t_1 + \log \beta t_4$ $+ 2 \log \alpha t_3$	$2 \log \beta t_1 + 2 \log \beta t_4$ $+ \log \alpha t_3$	$3 \log \beta t_1 + \log \beta t_4$ $+ 2 \log \beta t_5$	$T_2 > T_1 > T_3$
(0, 3)	$3 \log \beta t_1 + 3 \log \alpha t_3$	$3 \log \beta t_1 + 3 \log \beta t_4$	$3 \log \beta t_1 + 3 \log \beta t_5$	$T_1 > T_2 > T_3$

observed patterns, here 2 and any value between the two extremes. Note that  $n = 2$  is the bootstrap case: there is no more variability in the pattern than the one observed in the data. We will discuss both, starting with 256.

**256 possible patterns:** in the worst case, a pattern  $x$  has probability  $1/2$ , so that  $\theta^x(1 - \theta^x + \varepsilon) \simeq 1/4$ . Several patterns turn out to realize  $\|\log P^T - \log P^{T'}\|_\infty$ , all those of the form  $abab$  and  $aabb$  where  $a$  is any purine,  $b$  any pyrimidine or vice-versa. For these patterns,  $|\log \mathbb{P}(x; T) - \log P(x; T')| = \log(t_3/\beta t_1 t_4)$ . When branch lengths are of the same order of magnitude and for small  $\alpha t_i$  (typically of the order  $10^{-2}$ ),  $\|\log P^T - \log P^{T'}\|_\infty$  is no more than a few units, let us say 5. The troublesome case is  $\ell^T - \ell^{T'}$ . Since the empirical likelihood difference is  $\log(t_4/\alpha t_1 t_3)$ , we consider the true one no lower than a third of it. With all these

orders of magnitude, we can perform the calculation. The result is:

$$\begin{aligned}
\mathbb{P}(\ell_n^{T_2} > \ell_n^{T_1}) &\leq |\mathcal{A}|^s \times \max_{x \in \mathcal{A}^s} \exp - \frac{n(\ell^T - \ell^{T'})^2}{|\mathcal{A}|^{2s} \|\log P^T - \log P^{T'}\|_\infty^2 \theta^x (1 - \theta^x)} \\
&\leq 256 \exp \left( - \frac{3 \log^2 \left( \frac{t_4}{\alpha t_1 t_3} \right) / 9}{256^2 \log^2 \left( \frac{t_3}{\beta t_1 t_4} \right) / 4} \right) \\
&\leq 256
\end{aligned}$$

The result is disappointing but not unexpected. First, we have a low number of nucleotides so that our bound is extremely bad. Then, the number of possible patterns (512) is definitely too high and finally, it is doubtful that a single pattern  $x$  will a probability  $\theta^x$  close to  $1/2$ .

**2 possible patterns:** the only patterns are those observed in the data.  $\|\log P^T - \log P^{T'}\|_{app}$  is then  $\log(t_3/\beta t_1 t_4)$ . With the idea that  $T_1$  is better than  $T_2$ , pattern  $X_1$  is much more represented than pattern  $X_3$ , for example  $\theta^{X_1} \simeq 0.9$ . We then find  $\ell^{T_1} - \ell^{T_2} \simeq 0.9 \log(\alpha t_3/\beta t_4) + 0.1 \log(\beta t_1 t_4/t_3)$ . For  $t_3 = 2t_1 = 2t_4$ ,  $\alpha = 2\beta$  and  $\beta t_1 = 10^{-2}$ , we find:

$$\begin{aligned}
\mathbb{P}(\ell_n^{T_2} > \ell_n^{T_1}) &\leq |\mathcal{N}| \times \max_{x \in \mathcal{N}} \exp - \frac{n(\ell^{T_1} - \ell^{T_2})^2}{|\mathcal{N}|^2 \|\log P^T - \log P^{T'}\|_{app}^2 \theta^x (1 - \theta^x)} \\
&= 2 \times \exp \left( - \frac{3(0.9 \log 4 + 0.1 \log 0.005)^2}{2^2 \log^2(200) \frac{9}{100}} \right) \\
&= 2 \times \exp - \frac{29,02}{243,74} \\
&= 1.61
\end{aligned}$$

Since the probability is always lower than 1, this inequality seems disappointing. However it is only twice worse than the bootstrap bound, which considers the same amount of variability. It is also much better than the previous bound: the exponential decrease rate is  $-0.071$  compared to  $\simeq -5.10^{-6}$ . Of course, as  $n$  increases the inequality becomes sharper. For  $n = 50$  nucleotides, we obtain  $\mathbb{P}(\ell_n^{T_2} > \ell_n^{T_1}) \leq 0.57$  whereas its bootstrap counterpart is higher.

## A Proofs of the preliminary results

### A.1 Proof of the first result

We prove here Lemma. 6.

**Proof.** The proof is inspired by [10]. We note  $P_n^1 = \mathbb{P}(\sum_{i=1}^n (X_i - p) > \varepsilon)$  our probability of interest and  $Y_i = X_i - p$  the centered random variables. For any  $\lambda > 0$ , we have:

$$P_n^1 = \mathbb{P}\left(\sum_i Y_i > n\varepsilon\right) = \mathbb{P}\left(e^{\lambda \sum_i Y_i} > e^{\lambda n\varepsilon}\right) \leq \frac{\mathbb{E}[e^{\lambda \sum_i Y_i}]}{e^{\lambda n\varepsilon}} = \frac{\mathbb{E}[e^{\lambda Y_1}]^n}{e^{\lambda n\varepsilon}}$$

So that:

$$\begin{aligned} \log P_n^1 &\leq n \inf_{\lambda} \left( \log \mathbb{E}[e^{\lambda Y_1}] - \lambda \varepsilon \right) \\ &= n \inf_{\lambda} \left( \log (pe^{\lambda(1-p)} + (1-p)e^{-p\lambda}) - \lambda \varepsilon \right) \\ &= n \inf_{\lambda} \left( \log (pe^{\lambda} + (1-p)) - \lambda(p + \varepsilon) \right) \end{aligned} \quad (13)$$

The derivative of right-hand side of (13) with respect to  $\lambda$  is

$$-(p + \varepsilon) + \frac{pe^{\lambda}}{pe^{\lambda} + (1-p)}$$

so that  $\lambda^*$  the optimal value of  $\lambda$  is given by:

$$e^{\lambda^*} = 1 + \frac{\varepsilon}{p(1-p-\varepsilon)}$$

Replacing  $\lambda$  by  $\lambda^*$  in (13) gives:

$$\log P_n \leq n \left[ -(\varepsilon + p) \log \left( 1 + \frac{\varepsilon}{p(1-p-\varepsilon)} \right) + \log \left( 1 + \frac{\varepsilon}{1-p-\varepsilon} \right) \right] \quad (14)$$

The results is then obtained from applying the following lemma to (14). ■

**Lemma 11.** For all  $p \in (0, 1)$  and  $\varepsilon \leq p \frac{1-p}{1+p}$ ,

$$-(\varepsilon+p) \log \left( 1 + \frac{\varepsilon}{p(1-p-\varepsilon)} \right) + \log \left( 1 + \frac{\varepsilon}{1-p-\varepsilon} \right) \leq \frac{-\varepsilon^2}{2p(1-p-\varepsilon)} \left[ 1 - \frac{\varepsilon}{p(1-p-\varepsilon)^2} \right]$$

**Proof.** Note  $\Delta = \left[ -(\varepsilon + p) \log \left( 1 + \frac{\varepsilon}{p(1-p-\varepsilon)} \right) + \log \left( 1 + \frac{\varepsilon}{1-p-\varepsilon} \right) \right]$ . For  $x \in [0, 1[$ , it is well known that  $x - \frac{x^2}{2} \leq \log(1+x) \leq x - \frac{x^2}{2} + \frac{x^3}{3}$ . Using this for  $\varepsilon < p \frac{1-p}{1+p}$ , i.e.  $\frac{\varepsilon}{p(1-p-\varepsilon)} < 1$  we get the upper bound:

$$\begin{aligned} \Delta &\leq \left[ \frac{\varepsilon}{1-p-\varepsilon} - \frac{\varepsilon^2}{2(1-p-\varepsilon)^2} + \frac{\varepsilon^3}{3(1-p-\varepsilon)^3} \right. \\ &\quad \left. - \left( \frac{(p+\varepsilon)\varepsilon}{p(1-p-\varepsilon)} - \frac{(p+\varepsilon)\varepsilon^2}{2p^2(1-p-\varepsilon)^2} \right) \right] \\ &= \frac{-\varepsilon^2}{2p(1-p-\varepsilon)^2} [p + 2(1-p-\varepsilon) - 1] + \frac{\varepsilon^3}{3(1-p-\varepsilon)^3} + \frac{\varepsilon^3}{2p^2(1-p-\varepsilon)^2} \\ &= \frac{-\varepsilon^2}{2p(1-p-\varepsilon)} + \frac{\varepsilon^3}{2p(1-p-\varepsilon)^2} + \frac{\varepsilon^3}{3(1-p-\varepsilon)^3} + \frac{\varepsilon^3}{2p^2(1-p-\varepsilon)^2} \\ &= \frac{-\varepsilon^2}{2p(1-p-\varepsilon)} + \frac{\varepsilon^3 A(p, \varepsilon)}{6p^2(1-p-\varepsilon)^3} \end{aligned}$$

where  $A(p, \varepsilon) = 3 - 3p\varepsilon - 3\varepsilon - p^2$ . Maximizing  $A$  over  $\{(p, \varepsilon) / 0 \leq \varepsilon \leq p \leq 1\}$  gives  $\max A(p, \varepsilon) \leq 3$ , so that:

$$\Delta \leq \frac{-\varepsilon^2}{2p(1-p-\varepsilon)} \left[ 1 - \frac{\varepsilon}{p(1-p-\varepsilon)^2} \right]$$

For the lower bound, we get in the same way:

$$\begin{aligned} \Delta &\geq \left[ \frac{\varepsilon}{1-p-\varepsilon} - \frac{\varepsilon^2}{2(1-p-\varepsilon)^2} \right. \\ &\quad \left. - \left( \frac{(p+\varepsilon)\varepsilon}{p(1-p-\varepsilon)} - \frac{(p+\varepsilon)\varepsilon^2}{2p^2(1-p-\varepsilon)^2} + \frac{(p+\varepsilon)\varepsilon^3}{3p^3(1-p-\varepsilon)^3} \right) \right] \\ &= \frac{-\varepsilon^2}{2p(1-p-\varepsilon)^2} [p + 2(1-p-\varepsilon) - 1] + \frac{\varepsilon^3}{2p^2(1-p-\varepsilon)^2} - \frac{(p+\varepsilon)\varepsilon^3}{3p^3(1-p-\varepsilon)^3} \\ &= \frac{-\varepsilon^2}{2p(1-p-\varepsilon)} + \frac{\varepsilon^3}{2p(1-p-\varepsilon)^2} + \frac{\varepsilon^3}{2p^2(1-p-\varepsilon)^2} - \frac{(p+\varepsilon)\varepsilon^3}{3p^3(1-p-\varepsilon)^3} \\ &= \frac{-\varepsilon^2}{2p(1-p-\varepsilon)} + \frac{\varepsilon^3 B(p, \varepsilon)}{6p^3(1-p-\varepsilon)^3} \end{aligned}$$

where  $B(p, \varepsilon) = 3p^2(1-p-\varepsilon) + 3p(1-p-\varepsilon) - 2(p+\varepsilon)$ .  $B$  is decreasing in  $\varepsilon$  for all  $p \in [0, 1]$  so that  $B(p, \varepsilon) \geq B(p, 0) = 3p(1-p^2) - 2p$  which gives:

$$\frac{\varepsilon^3 B(p, \varepsilon)}{6p^3(1-p-\varepsilon)^3} \geq \frac{\varepsilon^3 B(p, 0)}{6p^3(1-p-\varepsilon)^3} = \frac{\varepsilon^3(3(1-p^2) - 2)}{6p^2(1-p-\varepsilon)^3} \geq \frac{-2\varepsilon^3}{6p^2(1-p-\varepsilon)^3}$$

And finally

$$\Delta \geq \frac{-\varepsilon^2}{2p(1-p-\varepsilon)} \left[ 1 + \frac{-2\varepsilon}{3p(1-p-\varepsilon)^2} \right]$$

■

## A.2 Proof of the second result

We prove in the same way Lemma. 7. The beginning of the proof is very similar to the previous one but the results differ.

**Proof.** We note  $P_n^2 = \mathbb{P}(\sum_{i=1}^n (X_i - p) > \varepsilon)$  our probability of interest and  $Y_i = p - X_i$  the centered random variables. For any  $\lambda > 0$ :

$$P_n^2 = \mathbb{P}\left(\sum_{i=1}^n (p - X_i) > \varepsilon\right) \leq \frac{\mathbb{E}[e^{\lambda Y_1}]^n}{e^{\lambda n \varepsilon}}$$

So that:

$$\frac{\log P_n^2}{n} \leq \inf_{\lambda} \left( \log (pe^{-\lambda} + (1-p)) + \lambda(p-\varepsilon) \right) \quad (15)$$

The derivative of the right-hand term of 15 with respect to  $\lambda$  is

$$(p-\varepsilon) - \frac{pe^{-\lambda}}{pe^{-\lambda} + (1-p)}$$

so that the optimal value  $\lambda^*$  of  $\lambda$  is given by:

$$e^{-\lambda^*} = 1 - \frac{\varepsilon}{p(1-p+\varepsilon)}$$

Replacing  $\lambda$  by  $\lambda^*$  in (15) gives:

$$\frac{\log P_n^2}{n} \leq \left[ -(p-\varepsilon) \log \left( 1 - \frac{\varepsilon}{p(1-p+\varepsilon)} \right) + \log \left( 1 - \frac{\varepsilon}{1-p+\varepsilon} \right) \right] \quad (16)$$

The result is then obtained from applying the following lemma to (16). ■

**Lemma 12.** *For every  $p \in (0, 1)$  and any  $\varepsilon \leq p$ , we have:*

$$-(p+\varepsilon) \log \left( 1 - \frac{\varepsilon}{p(1-p+\varepsilon)} \right) + \log \left( 1 - \frac{\varepsilon}{1-p+\varepsilon} \right) \leq \frac{-\varepsilon^2}{2p(1-p+\varepsilon)} \left[ 1 - \frac{2\varepsilon}{p(1-p+\varepsilon)} \right]$$

**Proof.** Note  $\tilde{\Delta} = -(p + \varepsilon) \log \left( 1 - \frac{\varepsilon}{p(1-p+\varepsilon)} \right) + \log \left( 1 - \frac{\varepsilon}{1-p+\varepsilon} \right)$ . For  $x \in [0, 1]$ , it is well-known that  $-\log(1-x) = \sum_{i=1}^{\infty} x^i/i$  and  $\log(1-x) \leq x - \frac{x^2}{2}$ . We can derive from these:

$$\begin{aligned}
\tilde{\Delta} &\leq (p - \varepsilon) \sum_{i=1}^{\infty} \frac{\varepsilon^i}{ip^i(1-p+\varepsilon)^i} - \frac{\varepsilon}{1-p+\varepsilon} - \frac{\varepsilon^2}{2(1-p+\varepsilon)^2} \\
&= \frac{-\varepsilon^2}{2p(1-p+\varepsilon)^2} [2(1-p+\varepsilon) - 1 + p] - \frac{\varepsilon^3}{2p^2(1-p+\varepsilon)^2} \\
&\quad + (p - \varepsilon) \sum_{i=3}^{\infty} \frac{\varepsilon^i}{ip^i(1-p+\varepsilon)^i} \\
&= \frac{-\varepsilon^2}{2p(1-p+\varepsilon)} - \frac{\varepsilon^3}{2p(1-p+\varepsilon)^2} - \frac{\varepsilon^3}{2p^2(1-p+\varepsilon)^2} \\
&\quad + (p - \varepsilon) \frac{\varepsilon^3}{p^3(1-p+\varepsilon)^3} \sum_{i=0}^{\infty} \frac{\varepsilon^i}{(i+3)p^i(1-p+\varepsilon)^i} \\
&= \frac{-\varepsilon^2}{2p(1-p+\varepsilon)} + \frac{\varepsilon^3 C(p, \varepsilon)}{2p^3(1-p+\varepsilon)^3}
\end{aligned}$$

where  $C(p, \varepsilon) = -p(1+p)(1-p+\varepsilon) + 2(p-\varepsilon) \sum_{i=0}^{\infty} \frac{\varepsilon^i}{(i+3)p^i(1-p+\varepsilon)^i}$ . Since  $0 < (1-p+\varepsilon) < 1$ ,  $\sum_{i=0}^{\infty} \frac{\varepsilon^i}{(i+3)p^i(1-p+\varepsilon)^i} \leq \sum_{i=0}^{\infty} \frac{\varepsilon^i}{p^i} = \frac{p}{p-\varepsilon}$  and  $C(p, \varepsilon) \leq -p(1+p)(1-p+\varepsilon) + 2p$ , so that

$$\frac{C(p, \varepsilon)}{2p^3(1-p+\varepsilon)^3} \leq \frac{2 - (1+p)(1-p+\varepsilon)}{2p^2(1-p+\varepsilon)^3} \leq \frac{2}{2p^2(1-p+\varepsilon)^3}$$

And we conclude:

$$\tilde{\Delta} \leq \frac{-\varepsilon^2}{2p(1-p+\varepsilon)} \left[ 1 - \frac{2\varepsilon}{p(1-p+\varepsilon)} \right]$$

■

## References

- [1] B. Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1):1–26, January 1979.
- [2] B. Efron, E. Halloran, and S. Holmes. Bootstrap confidence levels for phylogenetic trees. *Proc Natl Acad Sci U S A*, 93(14):7085–7090, Jul 1996.

- [3] J. Felsenstein. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39(4):783–791, July 1985.
- [4] J. Felsenstein and H. Kishino. Is there something wrong with the bootstrap on phylogenies? a reply to hillis and bull. *Systematic Biology*, 42(2):193–200, June 1993.
- [5] Joseph Felsenstein. *Inferring Phylogenies*. Sinauer Associates, September 2003.
- [6] Joseph Felsenstein. Statistical inference of phylogenies. *J.R. Statist. Soc.*, 146(3):246–272, 1983.
- [7] S. Holmes. Bootstrapping phylogenetic trees: Theory and methods. *Statistical Science*, 18:241–255, 2003.
- [8] T.H. Jukes and C.R. Cantor. *Evolution of protein molecules*, volume 3 of *mammalian Protein Metabolism*, chapter 24, pages 21–132. Academic Press, New York, 1969.
- [9] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, 16(2):111–120, Dec 1980.
- [10] P. Massart. *Concentration inequalities and model selection*. Springer, 2006.
- [11] M. Nei. Phylogenetic analysis in molecular evolutionary genetics. *Annu Rev Genet*, 30:371–403, 1996.
- [12] M.A. Newton. Bootstrapping phylogenies: Large deviations and dispersion effects. *Biometrika*, 83(2):315–358, 1996.
- [13] D. Penny and M. Hendy. Estimating the reliability of evolutionary trees. *Molecular Biology and Evolution*, 3:403–417, 1986.
- [14] H. Shimodaira and M. Hasegawa. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol*, 16(8):1114–1116, 1999.